CN 53-1189/P ISSN 1672-7673

# 一种天文卫星数据预处理方法*

赵海升[1,2]，葛明玉[1,2]，李正恒[1,2]，聂建胤[1,2]，宋黎明[1,2]
（1. 中国科学院高能物理研究所，北京 100049；2. 中国科学院粒子天体重点实验室，北京 100049）

**摘要**：天文卫星获取的数据需要经过一系列操作生成可以发布给科学用户的数据产品，数据预处理是其中比较重要的环节，它是数据发布的基础。以硬 X 射线调制望远镜卫星的中能 X 射线望远镜及天宫二号上伽玛暴偏振探测仪为例，详细介绍了数据的预处理步骤和方法。提出了一套数据预处理的步骤，并给出数据解析的策略和时间计算方法等，这对中国天文卫星数据处理系统的研制具有参考意义。

**关键词**：预处理；卫星数据；数据解包；时间计算
**中图分类号**：P171.5 **文献标识码**：A **文章编号**：1672-7673(2017)03-0376-06



卫星载荷获取的科学数据按预定义的格式在数据管理系统组包后发送给卫星平台，这些数据连同平台采集的数据通过数传通道下传至地面，经过简单的解码、去重复、分组等发送给卫星的科学系统。这些数据是采用国际数据系统咨询委员会[1]标准的数据包，它们要经过解包、转换、事例组装、时间计算、文件拆分和合并等处理步骤，生成可供进一步处理的数据单元，因为这里一般不涉及标定等处理，所以称之为数据预处理系统。不同卫星载荷的数据尽管在处理细节上不尽相同，但是大都需要这样的处理[2-4]，比如 Swift 卫星，下传的数据经过数据解包转换为 FITS 格式[5-6]，在数据分析中计算时间等。在数据分析之前，建立一套标准的处理流程，必将为其它卫星提供借鉴。本文主要以硬 X 射线调制望远镜卫星[7-8]的中能 X 射线望远镜(Medium Energy X-ray Telescope, ME)及天宫二号上的伽玛暴偏振探测仪[9]为例说明数据预处理系统的处理步骤及方法，希望给出一套标准的步骤和方法。

数据预处理中，根据数据的来源及功能，处理的数据分为：(1)科学事例数据(Event Files, EVTF)，主要是载荷记录的光子信息；(2)工程数据(Instrument Housekeeping Files, HKF)，载荷采集的温度、电压等信息；(3)平台数据(Platform Files, PLF)，卫星平台采集的轨道位置、姿态等。在预处理中还需要产生扩展的工程文件(EHK 文件, Extended Housekeeping Files, EHKF)，主要描述卫星的空间环境，比如进出南大西洋异常区信息、地磁截止刚度等，这些文件都是通过时间信息进行关联的。科学数据要求的时间精度比较高，时间来源也比较复杂，为了节省资源，一般只记录低位时间，其完整的时间需要地面通过算法给出，工程和平台数据的时间一般比较完整。另外，科学数据的存储不一定按时间顺序，尤其是来自不同电子学模块的数据，而工程和平台的数据一般按时间顺序存储。

数据预处理系统的任务：(1)对这些数据包进行校验，并进行数据拆分或者合并，解析包中的信息形成科学事例或者记录；(2)找出科学事例的完整时间信息，并对时间进行一定的修正，产生 EHK 文件；(3)产生可供进一步处理的完整的数据文件。图 1 是预处理系统的流程，输入的是经过简单处理的卫星数据，输出的是 1 级数据，1A、1B 和 1C 分别对应以上处理任务。

## 1 中能 X 射线望远镜和伽玛暴偏振探测仪介绍

中能 X 射线望远镜采用 Si-PIN 阵列，分成 9 个独立工作的探测单元，每个单元单独打包，然后

* 基金项目：国家自然科学基金（11503029）；中国科学院空间战略性科学先导科技专项（XDA04010300）资助.
收稿日期：2016-12-06；修订日期：2016-12-31
作者简介：赵海升，男，博士. 研究方向：天文数据分析及算法. Email：zhaohs@ihep.ac.cn

再循环读出到数据管理系统。事例只记录最低的14 位时间，高位时间通过进位事例记录，同时载荷将卫星接收的 GPS[10] 秒同步信号的到达时间作为一个特殊事例(GPS 秒脉冲事例，同时记录相对应的低位时间)插入数据流中，稍后对应的 GPS 秒同步时间，也作为一个特殊事例(GPS 秒同步事例)插入事例流中。

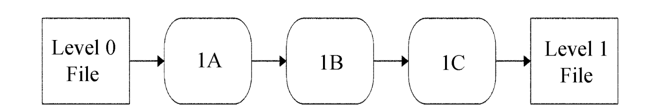



图 1　预处理系统的流程，其中 1A，1B 及 1C 为其中的处理步骤。Level 0，1 分别表示卫星下传的原始数据和产生的数据(可发布的数据)

Fig. 1　Data flow diagram. Level 0，1 represent the input telemetered data and the output data (can released to users) respectively

伽玛暴偏振探测仪由一个中心触发模块和25 个探测模块组成。中心触发模块记录触发时间(中心触发晶振提供，32 位时间码，记满清 0)及 25 个探测模块的触发标志，形成触发包，在触发的相应模块上，记录 23 位时间码(触发电路晶振提供，记满清 0)及触发通道，形成各个模块的独立数据包，因此伽玛暴偏振探测仪预处理的一个重点是将触发包与之相对应的触发模块产生的数据包进行匹配。它的 GPS 秒同步时间与中心触发晶振的对应关系保存于工程数据中，一分钟更新一次。

图 2 是中能 X 射线望远镜和伽玛暴偏振探测仪的数据处理流程。本文重点介绍它们的数据处理流程，并介绍其中的主要方法。

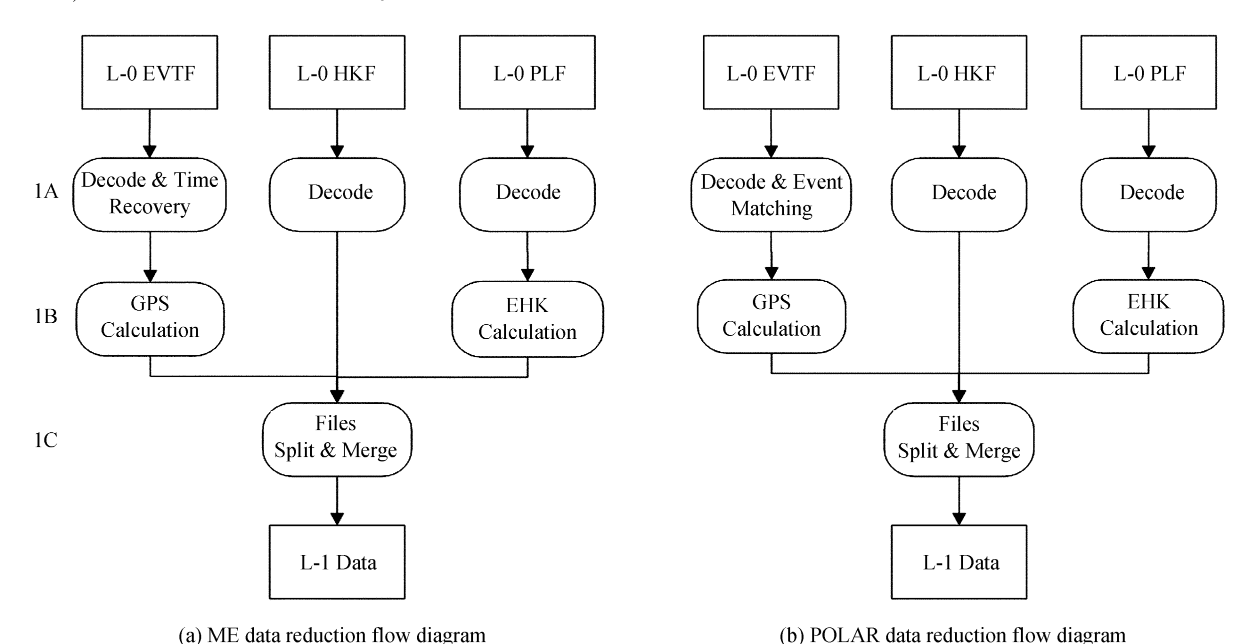



图 2　中能 X 射线望远镜和伽玛暴偏振探测仪的处理流程

Fig. 2　ME and POLAR data flow diagram

## 2　1A 级数据处理过程

该过程主要对所有数据进行解包，即将二进制数据按照格式解析为通用的十进制位数据，同时对科学数据进行一定的处理，比如中能 X 射线望远镜的本地时间计算(利用内部晶振表示时间，可以看作一个内部计数，计数与晶振周期之积表示时间长度)，伽玛暴偏振探测仪事例匹配等。

### 2.1　数据解析方法

一个数据包中可能包含多个不同的数据结构，比如科学事例包中，中能 X 射线望远镜包含 86 个相同的科学事例结构，而伽玛暴偏振探测仪包含 5 种不同科学数据结构(1 种触发包和 4 种事例包)。数据解析是将包中的所有数据结构解析。另外数据包头及其中数据结构的一些关键字也可以看作特殊的数据结构。这些结构的定义一般由载荷的制造者设计，软件人员按照设计逐一读出即可。本文采用配置文件解析的方法完成数据解析。配置文件的格式如表 1，ID 代表标识号，Title 为待解析数据的名

称，Type 为解析模式，它的取值：当为 0 时，如果 Repeat 为 1，则将 FB-EB 指定的区域进行解析，如果 Repeat 大于 1，则将 FB-EB 指定的区域分为 Repeat 份，分别进行解析；当为 1 时，将 FB-EB 指定的区域作为低字节位，将 F2B-E2B 指定的区域作为高字节位进行解析；当为 2 时，将 FB-EB 指定的区域进行解析，以 Step 为步长跳转到下一个区域(EB+Step+1，EB+Step+1+EB-FB)进行解析，以此类推。该模式下数组的维数可变。

所有模式下，起始位为数据结构的起始点。解析按照高位在前、低位在后的顺序进行，最大支持的解析位数为 64 位。

**表 1 配置文件格式说明**

**Table 1 The format of configure file for data decoding**

| Name | Contents |
| --- | --- |
| ID | Id number |
| Title | Title |
| Type | Decode pattern |
| FB/EB | Least significant word/byte/bit |
| F2B/E2B | Most significant word/byte/bit |
| Repeat/Step | Array/Step |
| Flag | Bit/Byte |

数据解析过程：读入上述配置文件，通过待解析数据的名称获得解析变量所在的索引，读入数据，根据索引找到变量所在的起始位置等属性。

2.2 解包策略

中能 X 射线望远镜和伽玛暴偏振探测仪的数据都按数据包组织，每一次只能解析一个包。处理的策略如下：

(1)检查包序号是否连续，对包进行循环冗余校验(Cyclic Redundancy Check，CRC)[11]；

(2)程序开始或者丢包发生时，寻找关键字，确定数据结构的开始位置，比如伽玛暴偏振探测仪的一个数据结构可能分到两个数据包中；

(3)对数据包进行组装，使之形成完整的数据包；

(4)读取一个完整的数据结构，对其进行循环冗余校验，利用该结构的配置文件进行解析，记录数据包序号、打包时间；

(5)对循环冗余校验不通过的均不进行解析而舍弃，对不完整数据结构也舍弃；

(6)数据的存储：对于科学事例要根据后续的处理功能选择是首先放入内存还是直接写入文件，对于工程数据和平台数据一般直接写入文件。数据的存储还要考虑方便数据文件的拆分和合并，比如中能 X 射线望远镜的 9 个探测器单元独立打包，在数据保存的时候采用独立保存的方法(可以保存为独立的文件，也可以保存为独立的数据区，比如在一个基于 ROOT 格式[12]文件中保存 9 个树形结构)。

2.3 中能 X 射线望远镜时间计算及伽玛暴偏振探测仪事例匹配

影响两者的主要因素是事例丢失及数据包丢失，它们都以特殊事例为参考点进行处理。

中能 X 射线望远镜的时间计算是指利用进位事例及事例的低位时间信息将事例的本地时间表示出来，因此，寻找事例的高位信息就成为重点。时间计算的参考事例为相邻进位事例，主要方法是扫描低位时间，记录翻转次数，如果翻转次数与相邻进位一致，中间进位时间则准确得到，如果不一致，则前后丢包的数据段不能给出进位时间。如果处理的数据仅有一个进位，则需要通过向后或者向前推断时间，如果一旦出现丢包，则后续事例不能得到准确的时间。

伽玛暴偏振探测仪事例匹配是将触发包与其对应的事例包进行匹配，匹配后触发包可看作一次物理过程。时间计算以基线触发包和基线事例包为参考点，方法是计算其它事例相对于参考点的时间差，如果触发包相对于基线触发包的时间差与事例包相对于基线事例包的时间差一致，且触发包中有该模块的触发标记，则认为该事例包属于触发包。匹配到新的基线时，则更新参考点以保证不受电子学翻转的影响。表 2 是一段已经匹配的结果，其中第 1 个记录为基线，也就是参考点。基线的确定方法为在事例流中触发包在前，事例包在后，先假定两者匹配，检查前后其它非基线包，如果匹配度(匹配上的事例与检查的所有事例之比)较高，则认为假定正确，否则继续寻找。

对工程数据和平台数据来说，1A 级数据处理过程最主要的是解包，一般不存在特殊处理，并按解包顺序依次存储；对科学数据来说，该过程包括解包、处理及事例分类存储，处理需要保证解析正确，在存储上需要考虑以后合并和拆分的要求，本文建议分单元存储。

表 2　一段触发包与事例包匹配的结果，其中第 1 个记录为基线信息，Flag 表示触发的探测模块

Table 2　Corresponding relationship between data packets and trigger packet. The first list is from pedestal packet and flag means the triggered modules in trigger packet

| Trigger packet | | | Data packet | |
|---|---|---|---|---|
| Local time (32bits) | High 23 bits | Trigger flag | Local time (23bits) | Module |
| 15562 43742 | 3039538 | 1-25 | 243995 | 1-25 |
| 15562 45471 | 3039541 | 8 | 243998 | 8 |
| 15562 78832 | 3039607 | 14 15 | 244063 | 14 15 |

## 3　1B 级数据处理过程

### 3.1　GPS 时间计算

该过程主要计算事例文件中事例的 GPS 时间，并将所有文件的一些解析量转换为实际物理量。GPS 时间的计算需要本地时间与 GPS 时间的一个对应关系(称之为对应点关系)，并且需要知道这个对应点使用的时间段(本地时间存在定期清零，该关系称之为对应段关系)，还需要知道晶振的最小时间周期，标称与实际可能存在差异，而且随着时间的推移及温度的变化，晶振频率也会发生变化。

中能 X 射线望远镜的时间计算中已经计算了 GPS 秒脉冲的本地时间，这可以看作 GPS 时间与本地时间的对应关系。伽玛暴偏振探测仪则将这个关系写入 HK 文件，每分钟更新。但是中能 X 射线望远镜和伽玛暴偏振探测仪的这种对应关系受晶振及 GPS 不稳定影响，如果进行一定的修正，效果更好。修正的方法是将这些对应点进行积累，分别计算这段时间的本地计数差 $\Delta N^{\mathrm{Local}}$ 和 GPS 时间差 $\Delta N^{\mathrm{GPS}}$，这样就很容易得到时钟周期 $t=\Delta N^{\mathrm{GPS}}/\Delta N^{\mathrm{Local}}$ 及平均对应点 $G+\Delta N^{\mathrm{GPS}}/2$，$N+\Delta N^{\mathrm{Local}}/2$，其中 $G$，$N$ 为初始的 GPS 起点和本地计数起点，利用这些信息可以比较好地计算事例的 GPS 时间 $T^{\mathrm{evt}} = G + \Delta N^{\mathrm{GPS}}/2 + (N^{\mathrm{evt}}-N-\Delta N^{\mathrm{Local}}/2) * t$，$T^{\mathrm{evt}}$，$N^{\mathrm{evt}}$ 分别表示事例的 GPS 时间和内部时间计数，具体积累的数据量与载荷的温度变化程度及要求的时间精度有关。

中能 X 射线望远镜的进位事例是一直累积的，原则上一个对应点可以适用于所有的数据，但是望远镜可能会重启，这样进位事例记录的晶振时间也会清零，而伽玛暴偏振探测仪的中心触发模块的时钟计数器会定时清零。处理的方法是：1A 级处理过程中记录本地时间周期翻转的开始和结束时间 $t_1$，$t_2$，此时间可以利用打包时间表示，打包时间来源于平台时间或者 GPS 时间，并对该时间段内的事例做标记，建立该时间段与事例之间的一一对应关系，即记录事例属于第几个时间段，因为内部时间翻转的位置是确定的，所以此关系唯一，这些时间段和事例都保存在文件中。1B 级处理过程中首先读入这些时间段，在每一个或者几个时间段内做时间对应点关系，对于中能 X 射线望远镜采用积累 GPS 事例的方法，而伽玛暴偏振探测仪则需要通过积累 HK 文件中 GPS 时间与内部晶振时间对应关系，进而得到对应点关系，再根据事例所属时间段标记进行 GPS 时间计算。

GPS 时间的误差为随机误差，短时间的稳定性不好，但是长期稳定性比较好，与之相反的是本地时间的短期稳定性好，长期存在时间漂移。当积累相当的 GPS 信号后，其误差趋于 0，这样比较容易得到晶振的周期，并可以长期刻画晶振的时间漂移。

1A 级过程解析的量均为整数，需要将其转换为实际对应的物理量或者将转换系数写入文件，本文建议采用后者的方式。

### 3.2　EHK 文件的产生

卫星下传的轨道和姿态数据不能完全表征卫星运行的空间环境，需要产生 EHK 文件。EHK 文件描述载荷运行的空间环境等，主要作用是将卫星运行的高本底区域、视场遮挡区域、载荷指向不稳定区域描述出来。这个文件按一定的时间间隔记录，主要内容包括：(1)卫星的位置、姿态及指向；(2)地球边界与视场的关系；(3)指向的晃动；(4)卫星所在的地磁截止刚度；(5)卫星与南大西洋异常区的关系；(6)卫星的指向与太阳、月球的关系等。这些量的具体定义可以参照各个科学卫星发布的数据产品[2-4]，需要说明的是，卫星的指向是对大量瞬时指向平均化的结果。EHK 文件产生依赖于

轨道文件、姿态文件、地磁截止刚度文件、南大西洋异常区文件以及跳秒文件等。

## 4 1C 级数据处理过程

这个过程主要对数据文件进行合并和拆分，并对数据进行一定的处理，比如排序、好时间段选择计算、坏探测器剔除等。中能 X 射线望远镜在 1A 级过程中产生的事例按独立探测模块分开，每一个独立单元的数据按时间排序，但是在此过程中所有单元的数据需要按时间顺序合并，可以利用合并排序算法，而伽玛暴偏振探测仪按触发包进行组织，没有排序过程。好时间段主要考虑载荷的关机时间或者降高压等时间、过南大西洋异常区时间等，一些损坏的探测器产生的数据需要剔除。

数据文件的获取采用数据库查询(比如 mysql 数据库[13])的方法，将 1A、1B 级产生的文件名称及属性写入数据库，其中属性包括了文件开始的时间及结束时间、文件类型(EVTF/HKF/EHKF 等)、级别(A/B)及路径等。给定一种数据类型及观测的起止时间即可从数据库中将所有的与此类型相符、且时间有交集的文件查询出来，然后对这些文件中的事例进行筛选、排序及处理。

这一过程中还可能包含数据格式的转换，比如提取 ROOT 的数据，在处理的同时转为天文分析常用的 FITS 格式，并将一些关键信息，比如文件的起始时间、观测位置等作为关键字[14]写入输出文件。

## 5 总 结

数据预处理是天文卫星数据处理环节中比较重要的一步，它是数据标定和数据分析的基础。不同载荷的数据处理细节也不尽相同，但是大都遵循同样的步骤，具体实现也有彼此参考的价值。本文在实现步骤上，载荷获取的科学数据经过解包和处理、GPS 时间计算以及文件的合并和拆分。载荷工程数据主要经过解包，平台数据则需要经过解包和计算，生成 EHK 文件，有些卫星考虑到 EHK 与其它文件的相关性，而将其产生步骤放到后续处理中。这些过程中，还要考虑物理量的转换及步骤之间的依存。1A 及 1B 文件原则上一一对应，这可以保证文件的追溯性，而 1C 的文件根据用户需要进行拆分和合并，比如需要一轨数据还是需要一天的数据，用户只需指定起止时间即可。实现细节上，数据解包采用配置文件方式描述数据结构，这一设计在软硬件之间建立桥梁，可以比较好地弥补设计初期的一些不稳定因素。数据处理中提出以特殊事例之间的数据为处理单元，处理中采用前向、后向分别处理，这尽管损失一些计算效率，但是提高了处理的正确性，并极大地使用数据信息，尽可能将所有事例考虑在内。GPS 计算中，利用 GPS 的长期稳定性，通过建立 GPS 时间与本地时间的对应点及对应段关系，从而利用这些关系计算 GPS 时间。数据存储采用了“总分总”的形式，比如中能 X 射线望远镜的科学数据源包是包括所有单元的，解包后按单元分开，文件合并时再对这些单元的数据合并，这样做是为了不同文件合并时，首先按单元合并，再利用排序将不同单元的数据合并。

本文没有涉及这些步骤的调度，特别是处理数据之间的依存关系，同时也没有涉及标定及以后的数据分析工作。

**致谢**：感谢中国科学院高能物理研究所 HXMT 地面系统及伽玛暴偏振探测仪组的帮助。

**参考文献**：

[1] Space Communications Protocol Specification-File Protocol. Recommendation for Space Data System Standards [R/OL]. (2014) [2016-12-06]. https://www.researchgate.net/publication/2593511_Recommendation_For_Space_Data_System_Standards.

[2] Suzaku Term. The Suzaku data reduction guide [OL]. (2013-08-01) [2016-12-06]. http://heasarc.gsfc.nasa.gov/docs/suzaku/analysis/abc/.

[3] RXTE Term. The ABC of RXTE [OL]. (1999-09-14) [2016-12-06]. http://heasarc.gsfc.nasa.gov/docs/xte/abc/contents.html.

[4] Capalbi M, Perri M, Saija B, et al. The swift XRT data Reduction Guide [OL]. (2005-04-01) [2016-12-06]. https://swift.gsfc.nasa.gov/analysis/xrt_swguide_v1_2.pdf.

[5] Wells D C, Greisen E W, Harten R H. FITS: A flexible image transport system [J]. Astronomy & Astrophysics Supplement, 1981, 44: 363-370.

[6] NASA. A primer on the FITS data format [OL]. (2014-10-28) [2016-12-06]. http://fits.gsfc.nasa.gov/fits primer.html.

[7] Li Ti-Pei, Wu Mei. The hard X-ray modulation telescope mission [J]. Physics, 2008, 37(9): 648-651.

[8] Li Ti-Pei. HXMT: A Chinese high-energy astrophysics mission [J]. Nuclear Physics B, 2007, 166: 131-139.

[9] Produit N, Barao F, Deluit S, et al. POLAR: a compact detector for gamma-ray bursts photon polarization measurements [C] // Nuclear Instruments and Methods in Physics Research, 2005, 550(3): 616-625.

[10] 蔺玉亭，赵东明，高为广，等. GPS 时间系统及其在时间比对中的应用 [J]. 地理空间信息，2009，7(3)：30-32.
Lin Yuting, Zhao Dongming, Gao Weiguang, et al. GPS time system and tts application to time comparing [J]. Geospatial Information, 2009, 7(3): 30-32.

[11] 韩家玮，张洪群. 基于 CCSDS 标准的遥感卫星下行数据纠错校验技术 [C] // 中国空间科学学会空间探测专业委员会学术会议，2007，29(1)：112-116.

[12] The ROOT Term. ROOT user's guide [OL]. (2014-05-01) [2016-12-06]. https://root.cern.ch/root/htmldoc/guides/users-guide/ROOTUsersGuide.html.

[13] 钟守波，韩波，张彦霞，等. 天文大数据管理工具的设计与实现 [J]. 天文研究与技术，2015，12(4)：510-517.
Zhong Shoubo, Han Bo, Zhang Yanxia, et al. Design and implementation of s software tool package for managing massive astronomical data [J]. Astronomical Research & Technology, 2015, 12(4): 510-517.

[14] 梅盈，刘东浩，王锋，等. 中国频谱射电日像仪 FITS-IDI 文件格式研究 [J]. 天文研究与技术——国家天文台台刊，2014，11(4)：388-395.
Mei Ying, Liu Donghao, Wang Feng, et al. A study of the FITS-IDI format for the Chinese Spectral Radio Heliograph [J]. Astronomical Research & Technology——Publications of National Astronomical Observatories of China, 2014, 11(4): 388-395.

## A Method of Data Preprocessing for Astronomical Satellite

Zhao Haisheng[1,2], Ge Mingyu[1,2], Li Zhengheng[1,2], Nie Jianyin[1,2], Song Liming[1,2]

(1. Institute of High Energy Physics, Chinese Academy of Sciences, Beijing 100049, China, Email: zhaohs@ihep.ac.cn;
2. Key Laboratory of Particle Astrophysics, Chinese Academy of Sciences, Beijing 100049, China)

**Abstract**: The data obtained from astronomical satellite is telemetered to ground and converted into high products (such as energy spectra and light curves) for further analysis. In one of these steps, data preprocessing is a very important step. This work will take the Media Energy X-ray Telescope (ME) on the Hard X-ray Modulation Telescope satellite and the Gamma-ray Burst Polarimeter-POLAR on Tiangong-2 (TG-2) as examples to introduce the method of data preprocessing. In this work, we present a standard data flow for astronomical satellite, and also introduce how to use configure file to decode data and how to calculate the local time of the payload and GPS time. This is helpful to establish data preprocessing system for other astronomical satellites.

**Key words**: Data preprocessing; Satellite data; Packet decoding; Time calculation